ML2020SPRING HW2 Report

學號：R08946015 系級：資料科學碩一 姓名：陳鈞廷

1. 請比較你實作的generative model、logistic regression 的準確率，何者較佳? 請解釋為何有這種情況？

   Generative model 與 logistic regression 的準確率分別如表格所示，logistic regression在 Training 與 validation 的準確率皆高於 generative model。
   我認為這次 Dataset 中的 feature 數量太多了，而且大部分的 feature 為 discrete，因此 generative model 較無法抽取到有用的 feature，造成其準確率不及 logistic regression。

	Training Set	Validation Set
Generative model	0.79470	0.79356
Logistic regression	0.86154	0.85886

2. 請實作 logistic regression 的正規化(regularization)，並討論其對於你的模型準確率的影響。接著嘗試對正規項使用不同的權重 (lambda)，並討論其影響。

   我將 logistic regression 的訓練次數 epoch 固定為 500，batch size 固定為 256， learning rate 固定為 0.01，每次更動 regularization 的 λ 值，以探討 regularization 的影響。
   從下方表格可以發現，當 λ 設定為 0.1 時，準確率明顯的下降了，而當 λ 設定在 0.001 以下時，準確率並沒有太大的變化，因此我認為在未實作正規化的情況下，模型並沒有 overfitting 的情形，因此選擇適當 λ 值時的正規化並不會對準確率有明顯的影響。

	Training Acc.	Validation Acc.
0.1	0.83958	0.83946
0.01	0.87648	0.87964
0.001	0.88448	0.88756
0.0001	0.88595	0.88793
0	0.88611	0.88775

3. 請說明你實作的best model，其訓練方式和準確率為何?

   我實作的 best model 是手刻的 logistic model。在 feature 的選擇上，使用 scikit-learn 提供的 GradientBoostingClassifier 來尋找權重較高的 feature，這裡選出了 199 個 feature，接著加上 continuous features (age, wage 等) 的 2 次方以及 3 次方項，最後總共有 213 個 feature。在 training 和 validation 的分割比例為 4: 1，training 和 validation set 中 positive sample 的比例皆為 0.20 左右。在模型的訓練上，經過多次調整，總共訓練了 500 個 `epoch`，`mini_batch_size` 為 256，`learning_rate` 調小至 0.01，並設定 `regularization lambda` 為 0.0001，用來抑制 overfitting 的現象。

	Public Score	Training Acc.	Validation Acc.
Logistic Regression	0.89276	0.88595	0.88793

4. 請實作輸入特徵標準化 (feature normalization) ，並比較是否應用此技巧，會對於你的模型有何影響。

   下方表格為實作特徵標準化前後的模型準確率比較，可以發現特徵標準化對於 generative model 沒有影響，而 logistic regression 的在實作 feature normalization 的準確率大約只提升了 0.001，究其原因應該是因為這個 data set 只有 7 個 continuous features ，而我在訓練模型時總共選了將近 200 個 feature，因此這些 continuous features 的 normalization 對於所有的 feature 來說影響非常小。

	Training Acc.	Validation Acc.
Generative model (with normalization)	0.79470	0.79356
Generative model (without normalization)	0.79470	0.79356
Logistic regression (with normalization)	0.88312	0.88517
Logistic regression (without normalization)	0.88190	0.88332